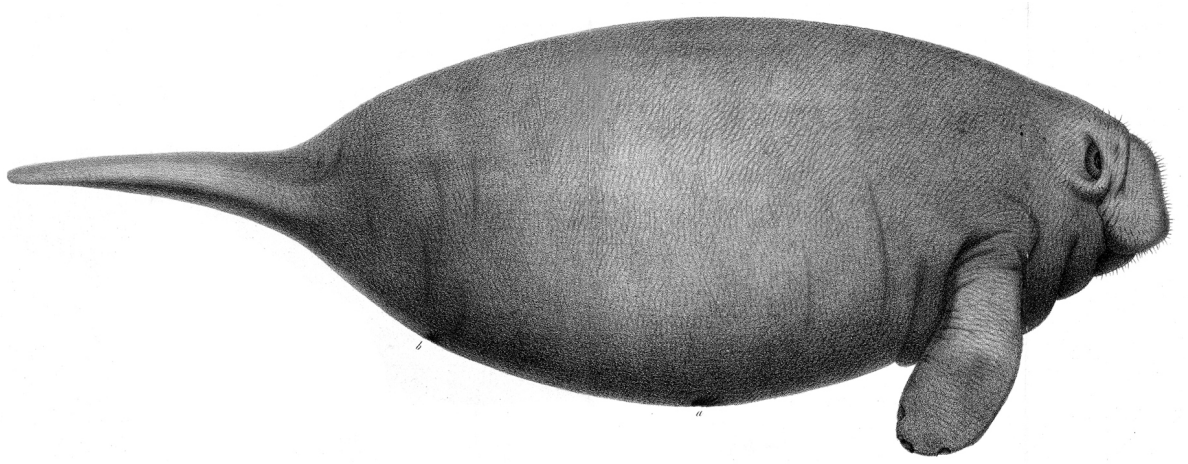
GBIF
www.gbif.org

Publishing and Registering Data with GBIF

Version 1.0



April 2011

Suggested citation:
GBIF (2011). Getting Started, Overview of Data Publishing in the GBIF Network, (contributed by Braak, K., Remsen, D., Hahn, A., Ko, B., Chavan, V., Raymond, M., Copenhagen: Global Biodiversity Information Facility, 16 pp. Accessible at http://links.gbif.org/dwc-a_publishing_guide_en_v1

ISBN: 87-92020-29-1

Persistent URI: http://links.gbif.org/dwc-a_publishing_guide_en_v1

Language: English

Copyright © Global Biodiversity Information Facility, 2011

License:

This document is licensed under a Creative Commons Attribution 3.0 Unported License

Document Control:

Version	Description	Date of release	Author(s)
1.0	Proofed for evaluation version	8 April 2011	DR

Cover: juvenile manatee, *Trichechus manatus*

This document is also part of the 'GBIF Data Publishing Manual version 1.0, ISBN 87-92020-31-3, available at http://links.gbif.org/data_publishing_manual

About GBIF

The Global Biodiversity Information Facility (GBIF) was established as a global mega-science initiative to address one of the great challenges of the 21st century – harnessing knowledge of the Earth's biological diversity. GBIF envisions 'a world in which biodiversity information is freely and universally available for science, society, and a sustainable future'. GBIF's mission is to be the foremost global resource for biodiversity information, and engender smart solutions for environmental and human well-being[1]. To achieve this mission, GBIF encourages a wide variety of data publishers across the globe to discover and publish data through its network.

.

[1] GBIF (2011). GBIF Strategic Plan 2012-16: Seizing the future. Copenhagen: Global Biodiversity Information Facility. 7pp. ISBN: 87-92020-18-6. Accessible at http://links.gbif.org/sp2012_2016.pdf

Table of Contents

About GBIF iii
Table of Contents iv
Introduction 1
Publishing data with GBIF 1
Using the Integrated Publishing Toolkit 1
Using Spreadsheet Processor or Make-Your-Own Darwin Core Archive (DwC-A) 2
Using TapirLink, BioCASe or DiGIR 2
Registering your dataset with GBIF 3
Using IPT 3
Using Spreadsheet Processor, Make-Your-Own DwC-A, or other community tools 3

Introduction

The purpose of this "Publishing and Registering Data with GBIF" guide is to instruct users to go through the publishing and registering steps after they have used tools to map their data or have produced a Darwin Core Archive file. A Frequently Asked Questions section is provided at the end to give users an estimate about the time required before their data are available on the GBIF Data Portal[2] and what will happen afterwards.

Publishing data with GBIF

After completing the previous steps, whether using the Integrated Publishing Toolkit (IPT), Spreadsheet Processor, Making your own DwC-A or a community option like TapirLink, the next step is to publish the dataset and its metadata, then register them with GBIF so the data are shared within the GBIF network.

To "publish" data means to make the data publicly available on the web. If the IPT, TapirLink, BioCASe Provider Software or DiGIR is used, the publishing functionality is built in. If the Spreadsheet Processor is used or a DwC-A file is generated manually, the user will need put the resulting files on a web server for download so they are "published".

Using the Integrated Publishing Toolkit

1. Follow the online user manual to get data uploaded and mapped[3].
2. In the "Managing Resources" page of your resource, there is a "Visibility" section with a "Public" button. Click the button and then your resource overview can be viewed from the home page of the IPT. At this stage, your resource is open to the public[4].
3. In the same "Managing Resources" page of your resource, there is a "Published Release" section with a "Publish" button. By clicking the "publish" button, IPT will generate a new release of your resource and make download links to a DwC-A file and an EML.xml, which contains the metadata. At this stage, your data is publicly available for download. Go to http://links.gbif.org/ipt_publish for more details.

[2] http://data.gbif.org/
[3] http://links.gbif.org/ipt_user_manual
[4] http://links.gbif.org/ipt_visibility

You must complete both step 2 and 3 to make your data ready to be registered.

Using Spreadsheet Processor or Make-Your-Own Darwin Core Archive (DwC-A)

If you selected to use Spreadsheet Processor or you have made your own DwC-A, you should have had a zip file as the product of the preparing procedure.

Now, to publish it, all you need to do is to put it in a web server that constantly connected to the Internet and serving contents to the public. For operations on a web server, it is strongly suggested to contact your web administrator to assist in making the file available online.

See Figure 1 for the file location on a typical web server running on Windows XP. Once you put the file within the "htdocs" folder, the file is then downloadable to the public. For example, if your web server is configured to have its domain name as "some.insitution.org," and the DwC-A file is named as "Specimen_NHM.zip,", once this zip file is put into the "htdocs" folder, people can download it by pointing their browsers to http://some.institution.org/Specimen_NHM.zip. You should keep this URL for registering with GBIF in the next step.

Please note that the IP address or the Domain Name of the web server should be stable, so the resource is constantly available online. Please also note that the URL is case sensitive

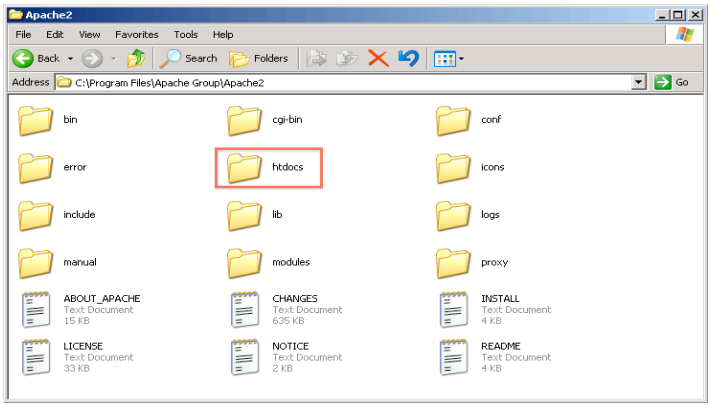


Figure 1. Web directory of Apache2 Web Server on Windows XP. Usually "htdocs," indicated by a red square, is the document root of the web server.

by default, and the access permission of the file should be properly set.

Using TapirLink[5], BioCASe[6] or DiGIR[7]

Please refer to the software guides for these tools to publish your data.

[5] TapirLink. http://wiki.tdwg.org/twiki/bin/view/TAPIR/TapirLinkManual
[6] BioCASe is a protocol designed by Biological Collection Access Services. See http://www.biocase.org/
[7] Distributed Generic Information Retrieval (DiGIR). See http://digir.sourceforge.net/

Once you have the metadata and dataset published using these tools, you should have both the metadata URL and dataset URL ready. Keep these URLs for the next step.

Registering your dataset with GBIF

Using IPT

The IPT supports automatic registration in the GBIF network. In the "Managing Resources" page of your resource, there is a "Visibility" section. If the status is public, then there should be a "Register" button and a drop-down list for institutions. Choose the institution with which the resource or dataset is associated, and click the "Register" button. Now your dataset and metadata are registered with the GBIFS. See the online manual for the IPT.

Using Spreadsheet Processor, Make-Your-Own DwC-A, or other community tools

There is no automatic registration for these options. Therefore, you need to use the data registration form at http://tools.gbif.org/dwca-register/ and provide the following information:

1. Dataset title
2. Dataset description
3. Technical contact (the person to be contacted in matters regarding technical availability or resource configuration issues on the side of the dataset or data publisher)
4. Administrative contact (the person to be contacted in all matters regarding scientific data content and usage of a specific dataset or data publisher)
5. Institution name
6. Your relation to this Institution
7. The name of the GBIF Participant Node that can endorse the publishing institution
8. The dataset URL: either the access point URL (if you are publishing using one of the provider softwares), or the DwC-Archive URL (if you are publishing via a zipped DwC-Archive)
9. The metadata document URL

Please ensure you have all of the information before you send the email. After sending the email, some questions are frequently asked:

How Long Until My Dataset Gets Registered?

Upon receiving your email, the GBIF Helpdesk will try to attend to your registration request as quickly as possible. The Helpdesk will first contact the GBIF Node selected with the registration, and ask them whether they want to endorse the new data provider installation in their domain. Each new registration needs to get formal endorsement from a GBIF Participant Node manager (who best knows the institutions and databases in their country/organisation) before it is allowed into the GBIF Registry. This is a simple quality control step required by the GBIF Participant Node Managers Committee.

Once endorsement has been received and the registration is completed, the registered dataset can be found on the GBIF Registry website[8]. Try searching by institution name or dataset title.

How Long Until My Dataset Gets Indexed By GBIF?

Following registration, the GBIF Helpdesk will queue the newly registered dataset for indexing. Depending on the size of the dataset, indexing can take anywhere from minutes to weeks. If problems are encountered during indexing, the GBIF Helpdesk will try to work with you to resolve them as quickly as possible.

When indexing is successful, the new dataset will become publicly available in the GBIF Data Portal (http://data.gbif.org) in the next "rollover" from the non-public indexing database with the public web portal database.

At present, GBIF attempts to update each registered dataset at least once every three months. The accumulated updates become public with each following "rollover", about every six weeks.

What Happens When My Dataset Is Indexed?

During indexing, a set of core data elements is retrieved from your dataset and is stored in the GBIF index, so that the dataset will become accessible for searches.

[8] http://gbrds.gbif.org